**Nourine Sabry – 211000198**

**ML assignment**

**Parkinsons’s disease prediction using machine learning**

**Project scope:**

Parkinson's disease (PD) is the second most common progressive neurodegenerative disease. PD symptoms include tremor, rigidity, cognitive impairment, and gastrointestinal issues. Diagnosing PD often relies on medical observations of motor symptoms, but in cases of early non-motor symptoms, subtle and mild symptoms might be overlooked. The aim of this project is to implement a machine learning model to try and predict whether a patient has PD or not based on speech features.

**Dataset overview:**

<https://www.kaggle.com/datasets/dipayanbiswas/parkinsons-disease-speech-signal-features/data>

This dataset consists of various speech signal features extracted from patients with PD and healthy individuals. The target variable ‘class’ indicates the presence (1) or absence (0) of PD. It consists of 756 rows and 755 columns. While it does not have duplicate values or missing data, it has highly correlated features and the data is also imbalanced.

**Steps:**

1. **Importing libraries & dataset:**

A screenshot of a computer

Description automatically generated

1. **Exploring the dataset & its features:**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

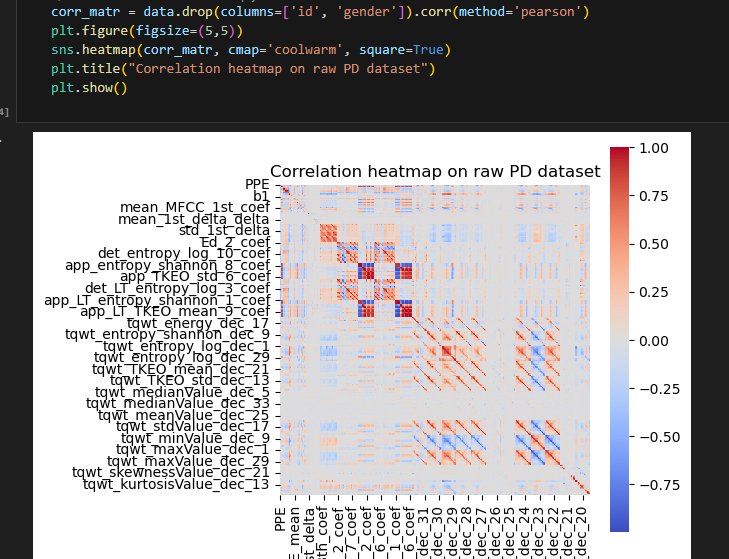
Description automatically generated**

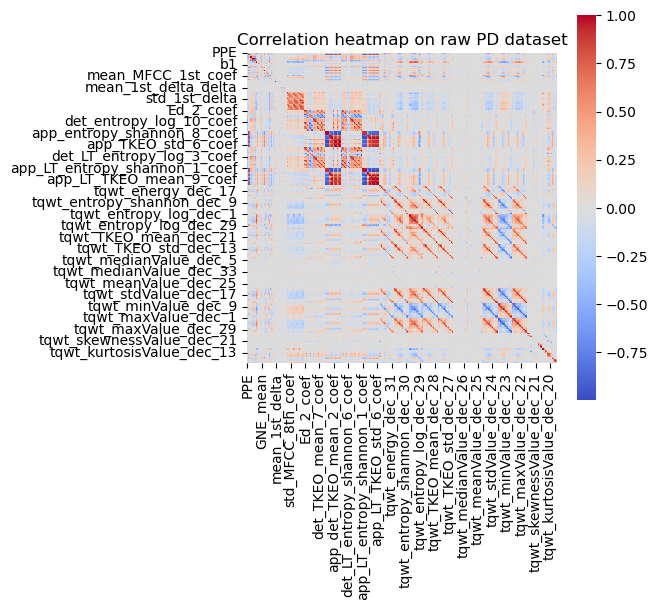
1. **EDA & data visualization:**

**A screenshot of a computer

Description automatically generated**

The above plot shows that in the raw dataset, females and males are underrepresented in the healthy group.





The above plot shows that many features are highly correlated.

A computer screen with text and numbers

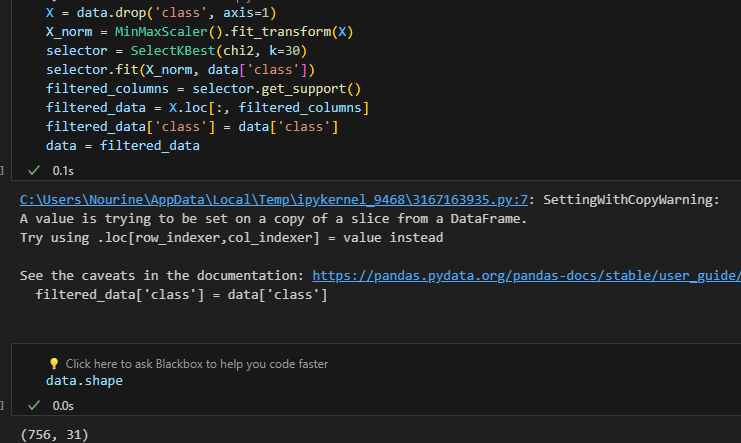
Description automatically generated

A blue and orange pie chart

Description automatically generated

The above plot shows that there is a data imbalance, the number of persons with PD is much higher than those who don’t have PD.

1. **Removing highly correlated features:**

****

Here, the feature space was reduced up from 755 to 30 using the chi square test, which is a text used to determine the level of similarity in the variances of the features. It is a widely used method for feature selection in machine learning due to its simplicity and computational efficiency.

1. **Balancing the data:**

A screen shot of a computer code

Description automatically generated

Here, the minority class (persons without PD) was upsampled.

1. **Separating features, splitting the dataset, & feature scaling:**

A screen shot of a computer code

Description automatically generated

1. **Training and comparing between two different models:**

* **Model 1: SVM RBF**

This model was chosen because it is highly effective in classification problems and tolerant to noise.

A screen shot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

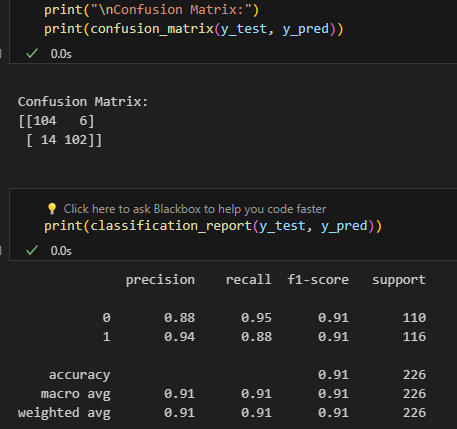
* **Model 2: Logistic regression**

This model was chosen because it is robust to irrelevant features, efficient, and well-suited for classification project.

A black screen with white text

Description automatically generatedA screen shot of a computer code

Description automatically generated



1. **Model deployment using streamlit**

**A screenshot of a computer

Description automatically generated**

**Results:**

While both SVM RBF and logistic regression performed good, logistic regression performed slightly better than SVM RBF, with 91.1% accuracy and 90.2% accuracy, respectively. Both models achieved a high F1 score, again, logistic regression has a slightly higher score than SVM RBF. Overall, logistic regression outperformed SVM RBF in all evaluation metrics used in this project.

**Limitations of this project:**

Two models were trained in this project, additional models could have been trained.

Requires domain knowledge in speech features.

**Conclusion:**

This is one of many in a large sea of machine learning projects that have shown the power of machine learning in solving real-world problems such as diagnostics, while no model is 100% accurate, machine learning has greatly impacted numerous fields such as healthcare, finance, and advertising. It has also shown that it is important to test multiple machine learning models, because it enables comparing between different models and seeing which one performs better than the others, which will help in deploying more accurate and efficient models. In this project, logistic regression slightly outperformed SVM RBF. This is probably due to the fact that logistic regression performs better on linearly separable data.